CLAIMS

Suy A12

1. A process for determining latency between multiple servers and a client across a network in a computer environment, comprising the steps of:

receiving a request for latency metrics on a content server;

wherein said latency metric request specifies a particular client;

providing a latency management table;

wherein said latency management table comprises a list of IP addresses along with corresponding Border Gateway Protocol (BGP) hop counts, dynamic hop counts, and Round Trip Times (RTT);

looking up the latency metric for said client in said latency management table;

sending said latency metric to the requesting server;

wherein the BGP hop count for said client in said latency management table is used for said latency metric upon the first request for said client; and

wherein the dynamic hop count and RTT data for said client in said latency management table are used for said latency metric for subsequent requests for said client.

2. The process of Claim 1, further comprising the steps of:

sending periodic latency probes to the IP addresses in said latency management table;

receiving response packets for said latency probes; and

recording the dynamic hop count and latency (RTT) data in said latency management table.

3. The process of Claim 2, wherein periodic latency probes are sent to a higher level server of a client by masking said client's IP address in said latency management table.

Sub A12

4. The process of Claim 1, further comprising the steps of:

receiving requests for a content server address from said client;

sending a latency metric request to the appropriate content servers;

receiving latency metric data from said content servers;

determining the optimal content server for said client; and

sending said optimal content server's address to said client.

5. The process of Claim 4, wherein said determining step gathers the expected latency metrics and uses the inverse relationship of the hop counts in said latency metric data in a weighted combination with the RTT in said latency metric data to determine which latency metric data indicates the optimal content server.

6. An apparatus for determining latency between multiple servers and a client across a network in a computer environment, comprising:

a module for receiving a request for latency metrics on a content server;

wherein said latency metric request specifies a particular client;

a latency management table;

wherein said latency management table comprises a list of IP addresses along with corresponding Border Gateway Protocol (BGP) hop counts, dynamic hop counts, and Round Trip Times (RTT);

a module for looking up the latency metric for said client in said latency management table;

a module for sending said latency metric to the requesting server;

wherein the BGP hop count for said client in said latency management table is used for said latency metric upon the first request for said client; and

wherein the dynamic hop count and RTT data for said client in said latency management table are used for said latency metric for subsequent requests for said client.

Sub A12 7

7. The apparatus of Claim 6, further comprising:

a module for sending periodic latency probes to the IP addresses in said latency management table;

a module for receiving response packets for said latency probes; and

a module for recording the dynamic hop count and latency (RTT) data in said latency management table.

8. The apparatus of Claim 7, wherein periodic latency probes are sent to a higher level server of a client by masking said client's IP address in said latency management table.

9. The apparatus of Claim 7, further comprising:

a module for receiving requests for a content server address from said client;

a module for sending a latency metric request to the appropriate content servers;

a module for receiving latency metric data from said content servers;

a module for determining the optimal content server for said client; and

a module for sending said optimal content server's address to said client.

10. The apparatus of Claim 9, wherein said determining module gathers the expected latency metrics and uses the inverse relationship of the hop counts in said latency metric data in a weighted combination with the RTT in said latency metric data to determine which latency metric data indicates the optimal content server.

11. A program storage medium readable by a computer, tangibly embodying a program of instructions executable by the computer to perform method steps for determining latency between multiple servers and a client across a network in a computer environment, comprising the steps of:

Sub A12

receiving a request for latency metrics on a content server;

wherein said latency metric request specifies a particular client;

providing a latency management table;

wherein said latency management table comprises a list of IP addresses along with corresponding Border Gateway Protocol (BGP) hop counts, dynamic hop counts, and Round Trip Times (RTT);

looking up the latency metric for said client in said latency management table;

sending said latency metric to the requesting server;

wherein the BGP hop count for said client in said latency management table is used for said latency metric upon the first request for said client; and

wherein the dynamic hop count and RTT data for said client in said latency management table are used for said latency metric for subsequent requests for said client.

12. The method of Claim 11, further comprising the steps of:

sending periodic latency probes to the IP addresses in said latency management table;

receiving response packets for said latency probes; and

recording the dynamic hop count and latency (RTT) data in said latency management table.

13. The method of Claim 12, wherein periodic latency probes are sent to a higher level server of a client by masking said client's IP address in said latency management table.

Sub A12

14. The method of Claim 11, further comprising the steps of:
receiving requests for a content server address from said client;
sending a latency metric request to the appropriate content servers;
receiving latency metric data from said content servers;
determining the optimal content server for said client; and
sending said optimal content server's address to said client.

15. The method of Claim 14, wherein said determining step gathers the expected latency metrics and uses the inverse relationship of the hop counts in said latency metric data in a weighted combination with the RTT in said latency metric data to determine which latency metric data indicates the optimal content server.